View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Xiamen University Institutional Repository

学校编码：10384
学号：200228032

分类号_____密级_____
UDC _____

厦 门 大 学
硕 士 学 位 论 文

异步分布式系统中故障检测器的设计与实现

Designing And Implementation of Failure Detector In Asynchronous Distributed Systems

王 良 明

指 导 教 师 ： 赵致琢 教授
专 业 名 称 ： 计算机应用技术
论文提交日期 ： 2005 年 6 月 日
论文答辩日期 ： 2005 年 7 月 日
学位授予日期 ： 2005 年 月 日

答辩委员会主席：_____
评 阅 人：_____

二〇〇五年六月

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果均在文中已明确标明。本人依法享有和承担由此论文而产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

摘 要

由于存在结点失灵的可能，Fischer 等人证明了异步系统中不存在一致合意的确定性求解算法[1]，为此，Sam Toueg 等人提出了异步分布式系统中的故障检测器技术[2]。故障检测器作为一个模块独立运行，并输出可疑结点列表，其他进程通过查询该列表来判断通讯对方结点的好坏。

以◇P 类[1]不可靠故障检测器为基础，采用模块化的体系结构，运用面向对象的编程语言，本文设计和实现了一个异步系统中的故障检测器 $FD=\{d_1,d_2,\cdots,d_n\}$，使得运行在第 i 台计算机上的故障检测服务模块 d_i 为本机上的所有结点提供故障检测服务，因此故障检测器 $FD=\{d_1,d_2,\cdots,d_n\}$ 也就能为整个网络上的结点提供检测服务。*FD* 把一个物理故障检测器映射为若干个逻辑故障检测器，逻辑故障检测器与本地结点一一“捆绑”在了一起，即 *FD* 对申请者来说是透明的。

FD 包含“结点接口”、“故障检测器”、“网络接口”和“结点和组群管理”四个模块，这四个模块相互协同工作，对外共同实现若干个逻辑故障检测器并“绑定”到本地的每一个应用结点（申请者）。

本文第三章对 *FD* 进行了全面的测试和性能分析，内容包括 *FD* 本身的正确性和健壮性（设计是否达到预期目标，能否在后台长时间无故障运行，各种数据表格维护是否正确，消息收发是否顺畅等）和 *FD* 性能分析（包括资源耗费情况，时间复杂性情况，消息复杂性和位复杂性情况，收敛快慢情况，发现结点失灵的响应时间情况等）。实验结果表明，本文设计的 *FD* 是实用的、可靠的和高效的。

关键词：分布式系统； *FD*；故障检测器；收敛性；失灵

[1] 具有强完全性（所有失灵结点最终被每个正确结点怀疑）和最终强精确性（在某个时刻之后，正确结点不被怀疑）属性的故障检测器统称为◇P 类。

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

Abstract

Fischer et al. had proved that there is no deterministic algorithm to solve consensus in asynchronous distributed systems, in which processes may crash. Hence Sam Toueg et al. had proposed unreliable failure detecttion technology. As a module, failure detector runs independently and outputs a list of suspected nodes. Other processes can access the list to judge whether the remote nodes are crrect or not.

Using Object-Oriented Language, we designed and implemented a Failure Detection Service, namely $FD=\{d_1,d_2,\cdots,d_n\}$, which has modular architecture and is based on unreliable failure detector class ◇P. The failure detection service module named d_i running on the ith host serves all of local nodes (subscribers) with failure detection service, it follows that *FD* serve all nodes in the distributed system with detection service. *FD* establishes a mapping of a physical failure detector to several logical failure detectors, and each of ones is associated with a subscriber. As a result, *FD* is transparent to subscribers.

FD consists of Object Interface module, Failure Detector module, Network Interface module and Object&Group Management module. Modules cooperate with each other and export logical failure detectors, which are associated with subscribers afterwards. The results of complete testing and performance analysis of *FD* are depicted in detail in the 3th chapter. Testing contents includes fds' correctnesss and robusticity itself (involving satisfying expectant objects or not, running correctly or not for ever in background, maintaining all kinds of data lists accurately or not, sending and receiving messages smoothly or not, etc.) and performance analysis of fds (involving resource consumption, time complexity, messages complexity, message bit complexity, the speed of convergence and the length of time to detect the crashed node, etc.). The data analysis in the experiment shows that this *FD* is practical and effective.

Key words: Distributed System; *FD*; Failure Detector; Convergence; Crashed

厦门大学博硕士论文摘要库

目 录

厦门大学博硕士论文摘要库

第一章 绪论

计算机与通信技术的结合催生了计算机网络，计算机网络和分布式算法的结合又推进了分布式应用的广泛开展。分布式应用正从数值计算密集型的专用领域（如石油矿藏定位、气象变化模拟和基因序列分析等）快速向非数值计算密集型的民用领域（如 Web 信息检索、分布式数据库等）扩展[3]。现实生活中绝大多数分布式系统属于异步系统，由于分布式系统的不确定性、通信延迟和可能的结点故障，要在异步系统上实现可靠的分布式应用，离不开故障检测器技术的支撑。下面分别对分布式系统和分布式算法以及故障检测器技术作一些简单介绍。

1.1 分布式系统和分布式算法

分布式系统是指互连在一起的具有自治能力的计算机、进程或处理器的集合。其中计算机、进程和处理器称为分布式系统中的结点（node），自治能力指明每个结点拥有自己独立的控制单元，“互连”表明结点间存在可以互相交换信息的链路。

为了方便信息交换和资源共享，人们发展了网络，之后，为了追求高可靠性和高性能的计算，人们又发展了分布式计算系统。与传统的集中式单机系统相比，分布式系统具有以下一些优点：

(1) 易扩充

根据负载的大小，很容易增减系统中的结点数，从而达到扩大或让出更多资源的目的。

(2) 更高的性价比

通常以廉价的低档 PC 机作结点组成分布式计算机系统，价格只有同等计算能力的大型机的几十分之一。

厦门大学博硕士论文摘要库

(3) 资源共享

各个结点的软硬件资源可被所有用户共享，从而极大地提高了资源的利用率。

(4) 更高的可靠性

因为控制、数据、软件和硬件具有多重性和分散性且可动态增减和重组的特点，使得分布式计算比集中式计算具有更高的可靠性。

(5) 更广泛的应用

集中式计算只是分布式计算的一个特例（只有一个结点的分布式系统），对于分散而又需要协调管理的行业（如银行、电信等），集中式计算不能满足用户的要求。

分布式系统对用户来说它的资源是透明的，用户只需要关心他的输入数据和运行结果，而不必关心系统在计算的过程中，具体用到哪些结点上的哪些资源。从这个意义上讲，分布式系统不同于计算机网络。分布式系统包含多个（可能是异构的）分散的、自治的处理资源，要使这些分散的结点很好地协调工作，面临一系列的挑战[3]：

(1) 资源的多重性使得差错处理和恢复问题变得很复杂，同时系统资源管理也变得异常困难，尽管全部资源发生故障的概率迅速下降。

(2) 资源的分散性使得计算的状态分散存放，因此分散的状态信息和不可预知的报文延迟使得系统的控制和同步问题变得很复杂，要想及时地、完整地搜集到系统各方面的信息并进行处理的最佳调度是很困难的。

(3) 系统资源的异构性意味着数据表示和编码、控制方式等均不相同，这样一来，就产生了翻译、命名、保护和共享等诸多问题。

判断这些问题是否可解、如何解以及解的复杂性则是分布式算法的任务。最初，术语“分布式算法”只指那些运行在地理上分布的许多计算机

厦门大学博硕士论文摘要库

上的协同算法，后来该术语的外延不断扩大，现在还包括了那些运行在局域网上的算法和共享存储器结构多处理机上的算法，而且还涵盖了大量的并发算法。

分布式算法就整体计算而言一般极为复杂，尽管算法的实际代码往往仅寥寥数行，因为运行这个代码的每一个进程向前推进的速度不一样，而且是动态变化的，这就意味着可能的全局状态空间异常庞大。所以，为了调试等目的而想重复过去的执行轨迹几乎是不可能的。相反，经典的集中式算法就显得简单而有序，从而也易于推导和证明。分布式算法与集中式算法的区别主要有以下几点：

(1) 没有全局状态

每个结点只拥有自己的局部状态，所有结点的局部状态之和组成全局状态，但这个全局状态实际上是不存在的。但集中式算法不同，由于只有一个进程，显然它自己的局部状态也就是全局状态，而这个全局状态是随时可得的。

(2) 缺少全局统一时钟

每个结点都有自己的局部时钟，时钟存在漂移，且漂移幅度又是不可预测的，同时各个结点执行的速度不一样，因此在分布式系统中企图统一时钟并保持同步是不可能的。Lamport 早在 1985 年就已证明网络分布式系统中不存在统一的全局物理时钟。相反，组成集中式算法的执行事件根据它们发生的时间排序，对于任意两个事件，它们不可能同时发生，因为执行是按顺序串行推进的，但两个分布式算法事件却有可能同时发生。

(3) 不确定性

同一个分布式算法的多次执行，结果很可能不一样的，这是因为每个进程的推进速度是变化的，由此导致由所有局部状态组成的全局状态是不

厦门大学博硕士论文摘要库

同的，自然结果也就不同。相反，集中式算法的计算结果只与输入相关，只要同一个算法的每次执行输入相同，自然结果一样。

设计分布式算法的难度要远远大于设计集中式算法的难度，设计者要不断变换角度并使思路跳跃腾挪。那么，怎么来判断一个新设计的算法好坏呢？通常采用消息复杂度、位复杂度和时间复杂度三个衡量指标。消息复杂度是指分布式算法在执行过程中传递消息的总次数，位复杂度是指传递的总比特数，而时间复杂度是指执行算法所需的时间单位数。在异步网络中，我们主要考虑消息复杂度和位复杂度，因为，在分布式系统中，消息传递延迟是影响算法性能最主要的因素。

在过去的 20 年里，分布式算法的研究主要围绕分布式基础算法展开，这些基础算法主要包括以下一些算法：

(1) 路由算法

目的是解决数据报文传递的路径选择问题。其中主要的技术手段是建立路由表和查询路由表，那么围绕着如何建表和查表，人们先后提出了最短路径路由、最少跳数路由、最少延迟路由、区间路由等算法。

(2) 波动算法

波动算法是一个满足以下三个属性的分布式算法：

A）终止性——算法的每一次运行都是有限的；

B）判定性——每次运行包括至少一个判定事件（一个特殊类型的内部事件）；

C）相关性——对于每次运行中的任何一个判定事件 *decide*，每个结点存在一个事件 *e*，*e* 在因果序上先于 *decide*。波动算法的一次运行称为一次波动。

(3) 选举算法

厦门大学博硕士论文摘要库

选举算法是一个波动算法，但还必须满足：

A）每个结点具有相同的局部算法；

B）算法是分散的，即结点的任意非空子集都能启动一次计算；

C）算法每次运行都能进入一个终止形态，且在每一个可达的终止形态中，只有一个结点处于领导人的状态，其它结点处于失败状态。

(4) 合意算法

通俗地讲，合意问题就是每个参与计算的正确结点提交一个值，最终所有正确结点一致认可（判定）其中的一个值。合意算法必须满足以下四个属性[2]：

A）终止性——每一个正确结点最终判定某个值；

B）一致性——不存在两个正确结点判定不同的值；

C）一致完整性——每一个正确结点至多判定一次；

D）一致有效性——如果一个结点判定v，那么之前v一定是被某个结点提交的。

(5) 终止检测算法

终止检测算法又称控制算法，由它监控的算法称为基本算法。控制算法由终止检测算法和终止发布算法组成，它必须满足：A）不能干扰基本算法的运行；B）如果基本算法隐式终止，那么控制算法必须在有限步内启动终止发布算法；C）终止发布算法启动时，基本算法必须已进入隐式终止状态。

(6) 快照算法

快照就是分布式系统的一个全局状态，可达的全局状态称为可行的。快照算法目的是用来生成可行的快照，可行的快照可用于计算恢复和死锁检测等。

厦门大学博硕士论文摘要库

1.2 故障检测器概述

在完全异步系统中不存在哪怕只有一个节点失灵（crashed）的合意问题求解算法[1]，因为我们无法判断一个节点到底是失灵还是运行速度非常慢。因此，为了在不可靠完全异步系统（存在失灵节点、消息丢失和通道失灵）中实现合意问题求解算法，人们先后提出了概率算法、放宽条件从而弱化问题或部分同步技术，但所有这些都是被动的。SAM TOUEG 等人另辟西径，于 1991 年发表了一篇基础性论文[2]，在文中作者提出了可靠分布式系统中的不可靠故障检测器模型。该模型确定故障检测器作为模块独立运行，它负责维护一个可疑节点列表，其他进程通过查询该列表来判断节点好坏。由于故障检测器普遍采用超时技术，由此招致一些人的怀疑和批评，他们认为超时技术本身就可以直接用来求解合意问题，但有意思的是故障检测器技术在责难声中还是得到了广泛应用。1997 年 Marcos 等人提出不用超时技术的心跳故障检测器[4]，该检测器输出以整数计数器为元素的向量，而不是可疑节点列表，其他进程通过观察计数器是否变化（比较计数器在两个不同时刻的值）从而间接判断相应节点好坏，如 Sam Toueg 等人给出的带心跳故障检测器的广播算法[4]就是通过简单比较两次取（查询故障检测器）值是否变化来判断节点好坏，显然这样的方法容易产生较高的误判率。

1.5 本文的主要研究内容

分布式基础算法的具体设计、实现和运用具有特殊性和多样性，在不同的应用场合人们往往根据具体的环境和目标采用不同的设计思想和实现方法，因而实际运用的性能也千差万别。本文通过具体设计和实现一个

厦门大学博硕士论文摘要库

◇P 类（具有以下两个属性：i）强完全性：所有失灵结点最终被每个正确结点怀疑；ii)最终强精确性：在某个时刻之后，正确结点不被怀疑）故障检测器并从体系结构、模块描述和性能评估等方面进行了深入地探讨。

由 n 台计算机组成的异步网络系统中存在 m 个分布式应用，这种情况在实际生活中普遍存在。最典型的一个例子就来自银行部门对数据的处理。在银行计算机网络上通常运行资金转帐、效益实时分析和安全监控等分布式应用。现在的问题是如何设计和实现一个故障检测器来高效低耗地检测这 m 个分布式应用中的故障结点？本文的目的就是回答这个问题，具体内容见第二章：**异步系统中故障检测器的设计**和第三章：**故障检测器的实现和实验分析。**

厦门大学博硕士论文摘要库

第二章 异步系统中故障检测器的设计

2.1 提出问题

结点会发生故障、消息传递存在任意延迟和任务调度不可预知，这是在异步分布式系统中设计和构建可靠分布式应用首先面临的事实，必须加以解决。例如，从 2000 年初发布的容错 CORBA 研究文档[5]中我们可以清楚地看到，定义故障管理系统成了容错 CORBA 工作组的一个主要目标。

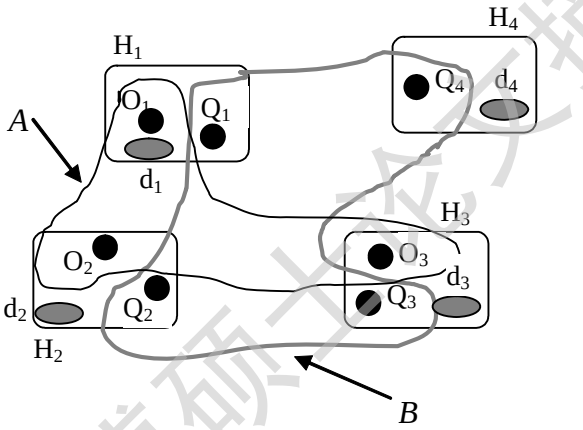

图 2.1 四台计算机上运行两个分布式应用

由 n 台计算机组成的异步网络系统中存在 m 个分布式应用，每个分布式应用包含不同个数的结点，例如在图 2.1 中，四台机器$\{H_1,H_2,H_3,H_4\}$组成一个异步网络系统，在这个系统之上同时计算两个分布式应用A和B，结点集合$\{O_1,O_2,O_3\}$组成分布式应用A，结点集合$\{Q_1,Q_2,Q_3,Q_4\}$组成分布式应用B。现在的问题是如何在多个分布式应用并存的环境下设计一个高效实用的故障检测器，由这个故障检测器负责所有结点的故障检测任务。

为此，本章以具有最终精确性属性的不可靠故障检测器为基础，深入

厦门大学博硕士论文摘要库

Degree papers are in the “Xiamen University Electronic Theses and Dissertations Database”. Full texts are available in the following ways:
1. If your library is a CALIS member libraries, please log on http://etd.calis.edu.cn/ and submit requests online, or consult the interlibrary loan department in your library.
2. For users of non-CALIS member libraries, please mail to etd@xmu.edu.cn for delivery details.

厦门大学博硕士论文摘要库